Synopsis of the thesis entitled

GENE EXPRESSION ANALYSIS OF TYPE-2 DIABETES WITH PARENTAL HISTORY – A COMPUTATIONAL APPROACH

Submitted for the award of the degree of

DOCTOR OF PHILOSOPHY

IN
COMPUTER SCIENCE AND SYSTEMS ENGINEERING

BY

V.CHANDRA SEKHAR

Under the guidance of

Prof. P.SRINIVASA RAO

Head of the department

Department of Computer Science and Systems Engineering

Andhra University College of Engineering (Autonomous)

DEPARTMENT OF COMPUTER SCIENCE AND SYSTEMS ENGINEERING
COLLEGE OF ENGINEERING(AUTONOMOUS), ANDHRA UNIVERSITY
VISAKHAPATNAM – 530 003, ANDHRA PRADESH, INDIA
JANUARY- 2013

INDEX

S.No	TOPIC	PAGE No.
1	INTRODUCTION ……	1
2	LITERATURE REVIEW…….....	2
3	PROBLEM STATEMENT ……	5
4	METHODOLOGY……	5
5	IDENTIFICATION OF DIFFERENTIAL GENES …	6
6	CLASSIFICATION OF IDENTIFIED GENES ……..	9
7	CONCLUSIONS …….	12
8	ORGANISATION OF THE THESIS……...	13
9	REFERENCES ……	13
10	PUBLICATIONS FROM THE WORK ……..	16

SYNOPSIS

1. INTRODUCTION

The prevalence of chronic diseases is increasing at an alarming rate. An epidemic of type 2 diabetes mellitus (T2DM) is sweeping across the world and the prevalence is projected to rise for several decades into the future. Factors that are associated with this rising frequency include excess adiposity and a variety of metabolic factors. Type 2 diabetes mellitus (T2DM) is a complex disease that represents a major public health concern around the world. Although we already know that alteration of the environmental and lifestyle risk factors can substantially reduce progression of this disease, the prevalence of diabetes is increasing every year. T2DM is frequently not diagnosed until complications appear because the effectiveness of early diagnosis through screening of asymptomatic individuals has not been established. Genes play an important role in the development of diabetes mellitus. Type 2 diabetes is a polygenic disorder with multiple genes located on different chromosomes contributing to its susceptibility. Although genetics could play an important role in the higher prevalence of this disease, it is not clear how genetic factors interact with environmental and dietary factors to increase their incidence. It is hoped that better understanding of the genes and gene expression analysis would help to identify potential genes causing Type-2 diabetes. Microarray experiments allow description of genome-wide expression changes in health and disease.

A gene is a unit of heredity in a living organism. To understand a genome more comprehensively, we need to move beyond the static view and understand how genes interact with each other. DNA microarrays, also known as DNA chips or gene chips, enable to measure thousands of genes simultaneously. The microarray is a multiplex lab-on a-chip. It is a 2D array on a solid substrate that assays large amounts of biological material using high-throughput screening methods. Microarray helps in estimating the amount of protein in the cell and a lot of information can be derived from this technology. Hence, microarrays provide a tool for answering a wide range of questions about the dynamics of cells. DNA microarray technology is used for two major applications: one to identify the sequence of gene and two to determine the expression level of genes.

2. LITERATURE REVIEW

Several methods are available in the literature to perform Differential Gene Expression Analysis to find the potential Genes causing various Diseases. A number of methods can be used to normalize microarray data and provide estimates of changes in gene expression that are corrected for potential confounding effects. This approach establishes a frame work for the general analysis and interpretation of micro array data.

Depending upon the kind of immobilized sample used to construct arrays and the information etched, the Microarray experiments can be categorized in three ways: Microarray expression analysis, Microarray for mutation analysis, Comparative Genomic Hybridization. In Microarray expression analysis, the cDNA derived from the mRNA of known genes is immobilized. The sample has genes from both the normal as well as the diseased tissues. Spots with more intensity are obtained for diseased tissue gene if the gene is over expressed in the diseased condition. This expression pattern is then compared to the expression pattern of a gene responsible for a disease. In the present work Microarray data analysis is used for identifying and classifying genes causing Type 2 Diabetes Mellitus (T2DM) with respect to parental history.

Type 2 diabetes is a lifelong (chronic) disease in which there are high levels of sugar (glucose) in the blood. Type 2 diabetes is the most common form of diabetes. It is well established that the prevalence of type 2 diabetes (T2DM) is rising worldwide [MAY, 2006] While environmental factors, such as obesity and lack of physical activity, play an important role to the rapid increase in the prevalence of T2DM, genetic factors are also important for the increase risk of T2DM [ATH, 2009]. Always there is a doubt why some people develop type 2 diabetes and others don't. It's clear that certain factors increase the risk, however, including Weight, Fat distribution, Inactivity, Family History, Race, Age, Pre-diabetes, Gestational diabetes. Obese adolescents with T2DM have hippocampal as well as prefrontal volume reductions relative to carefully matched non-insulin resistant obese adolescents [HAN, 2011].

Different forms of emotional stress are associated with an increased risk of the development of Type-2 diabetes, particularly depression, general emotional stress, anxiety,

anger/hostility and sleeping problems [FRA, 2010]. Weight gain in early adulthood is related to a higher risk and earlier onset of type 2 diabetes than is weight gain between 40 and 55 years of age [ANJ, 2006]. The impact of family history of diabetes on insulin dynamics has been confirmed in cross-sectional studies in adults [KLE, 1996] but not in younger children suggesting that the emergence of risk occurs at some point during growth and development [LOU, 2007]. The prevalence of diabetes in mothers was three fold higher than in fathers of T2DM patients [ATH, 2009].

Microarray is a high-throughput technology allowing the simultaneous screening of the expression levels of thousands of genes in one experiment. In microarray studies, for instance, a small fraction of genes typically exhibit significant differential expression among tens of thousands of genes whose expression levels are measured simultaneously. Thus, it is of great importance to identify genes relevant to a biological phenomenon of interest and to characterize their expression profiles [SAT and YAS, 2009].

In microarrays, the process of removing non-biological variation that is masking meaningful information is known as *normalization*. The correction of the data according to those factors, introducing either systematic or random errors, is an essential stage prior to the analysis and biological interpretation of the data [FAT, 2004]. Various normalization methods have been proposed [EFR, 2000; KER, 2001; SCH, 2000; SPE,1998; WOL, 2001] to reduce some of the variability. Chen et. al [CHE, 1997] considered normalization methods in terms of the ratio of fluorescence intensities within each array. They used the loess fit to adjust for intensity and location dependency biases. Kerr and Churchill [KER and CHU, 2001] recommended that an analysis should use all the information in the data and not reduce to ratios. They proposed an analysis of variance (ANOVA) model for individual red and green intensities. The ANOVA model simultaneously adjusts for the dye, within- and among-array effects globally. The ANOVA model uses the mean to estimate normalization factors. Delongchamp et. al [Del, 2002] recommended the median estimate since the median is more robust against the highly over- or under expressed genes.

There are three major types of applications of DNA Microarrays. The first involves finding differences in expression levels between predefined groups of samples (‘‘class

comparison’’). A second application, ‘‘class prediction,’’ involves identifying the class membership of a sample based on its gene expression profile (class prediction) . The third type of application involves analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features (class discovery) [ADI ,2006] .

A key goal is to extract the fundamental patterns of gene expression inherent in the data. Many mathematical techniques have been developed for identifying underlying patterns in complex data. Clustering techniques are one such methods for interpreting gene expression. Pablo Tamayo et. al [PAB, 1999] described the application of self-organizing maps, a type of mathematical cluster analysis that is particularly well suited for recognizing and classifying features in complex, multidimensional data. GENECLUSTER was used to organize the genes into biologically relevant clusters that suggest novel hypotheses about hematopoietic differentiation.

Haifeng Li et. al [HAI, 2004] proposed the minimum entropy criterion for clustering gene expression data. The experimental results show that the method performs significantly better than k-means/medians, hierarchical clustering, SOM, and EM, especially when the number of clusters is incorrectly specified. Xiaohua Hu and IllhoiYoo [XIA and ILL, 2004] presented a cluster ensemble framework for gene expression analysis to generate high quality and robust clustering results.

Type 2 diabetes mellitus (T2DM) is a complex disease that represents a major public health concern around the world. Although we already know that alteration of the environmental and lifestyle risk factors can substantially reduce progression of this disease, the prevalence of diabetes is increasing every year. T2DM is frequently not diagnosed until complications appear because the effectiveness of early diagnosis through screening of asymptomatic individuals has not been established. It is hoped that better understanding of the genes and micro array gene expression of the disease would help to improve treatment and prevention.

3. PROBLEM STATEMENT

Type 2 diabetes is a chronic disease in which there are high levels of sugar (glucose) in the blood. Type 2 diabetes is the most common form of diabetes. Genetic, environmental, and metabolic risk factors are interrelated and contribute to the development of type 2 diabetes mellitus. A strong family history of diabetes mellitus, age, obesity, and physical inactivity identify those individuals at highest risk. Current interventions for the prevention and retardation of type 2 diabetes mellitus are those targeted towards modifying environmental risk factors such as reducing obesity and promoting physical activity. Obesity and family history of diabetes are major predictors of type 2 diabetes. Both factors are relatively easy to assess and are widely used for the identification of individuals with undiagnosed diabetes. A parental history of diabetes is believed to reflect genetic susceptibility to hyperglycemia. Presence of a parental history is associated with impairments in insulin sensitivity and/or insulin secretion.

Various efforts are made in identifying the differential genes using gene expression analysis to identify the potential genes causing various disease like cancer, diabetes etc. A key goal is to extract the fundamental patterns of gene expression inherent in the data. Many mathematical techniques have been developed for identifying underlying patterns in complex data. Various Clustering techniques are proposed for interpreting gene expression. It is necessary to perform analysis for causing genes of Type 2 Diabetes based on parental history. The present work is aimed at addressing these issues in general with a specific focus on parental history.

4. METHODOLOGY

Gene expression analysis is a two step process that involves identifying differentially expressed genes and functionally classifying these differentially expressed genes. Appropriate statistical techniques are required to furnish realistic information on the differentially expressed (DE) genes. Outliers are often suspected as possible candidates for differential expression genes. By using the statistical methods like Mahalanobis Distance (MD) and Minimum Covariance Determinant, outliers' detection was performed. Upon identifying the outliers the next task is to

identify those outlier genes that are differentially expressed across the different samples. To know the biological significance of differentially expressed genes, functional classification has to be performed using Gene Ontology (GO). To determine pathways associated with differentially expressed genes, pathway analysis was performed. Prior to analysis, the data for each combination was normalized using Loess normalization.

5. IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

In any micro array study the primary objective is to assess miRNA transcript levels of samples under different experimental conditions. Which of the thousands of genes show significant difference in expression levels in the samples is the question of importance. Appropriate statistical techniques are required to furnish the accurate information on differentially expressed genes if there are no or limited replicates due to practical constraints in majority of the experiments. Data from samples were hybridized on Human 40 K OchiChip Array. Gene expression values were obtained after quantification of TIFF images. Empty spots and control probes were removed before proceeding with data analysis.

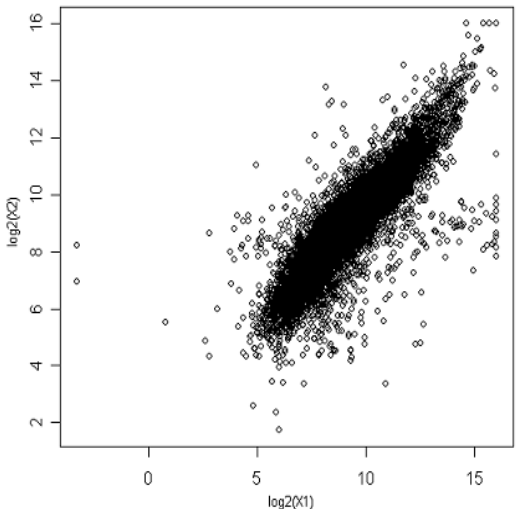


Fig 5.1 Scatter plot of log intensities of gene expression values for sample X1 and X2

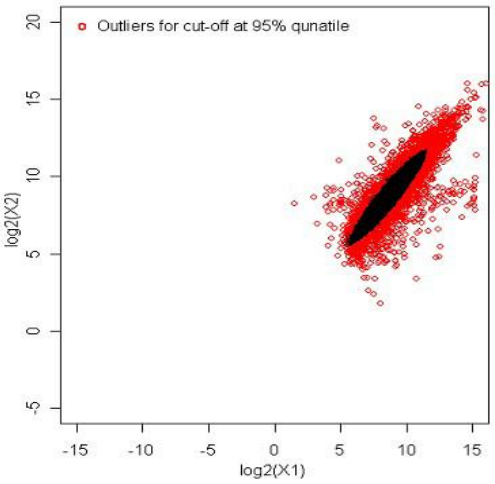


Fig 5.2 Scatter plot showing outliers at 95% cut-off of RD

For experiments with two samples ,the assumption is that the log intensity values of gene expression for the two samples (Fig. 5.1) are linearly related, following bivariate normal distribution, contaminated with outliers . In a contaminated bivariate distribution, the main body of the data is characterized by bivariate normal distribution and constitutes regular observations. The non-regular observations, described as outliers (Fig. 5.2), represent systematic deviations. These outliers are often suspected as possible candidates for differential expression genes. In the current study, an exploratory approach consisting of two-stages to detect outliers from bivariate population and determining differentially expressed candidates from these outliers. The approach provides the fold-change value considering the scatter of observations and thereby provide sup and down regulated genes across the samples (Fig. 5.4).

Stage- I: Multivariate Outlier Detection:

Outlier detection is one of the important tasks in any data analysis, which describe abnormalities in the data. Many methods have been proposed in the literature for detecting univariate outliers based on robust estimation of location and scale parameters. The standard method for multivariate outlier detection involves robust estimation of parameters in the *Mahalanobis Distance* (MD) measure and then comparing MD with the critical value of c2 distribution. The values larger than the critical value are treated as outliers of the distribution (Fig 5.3).

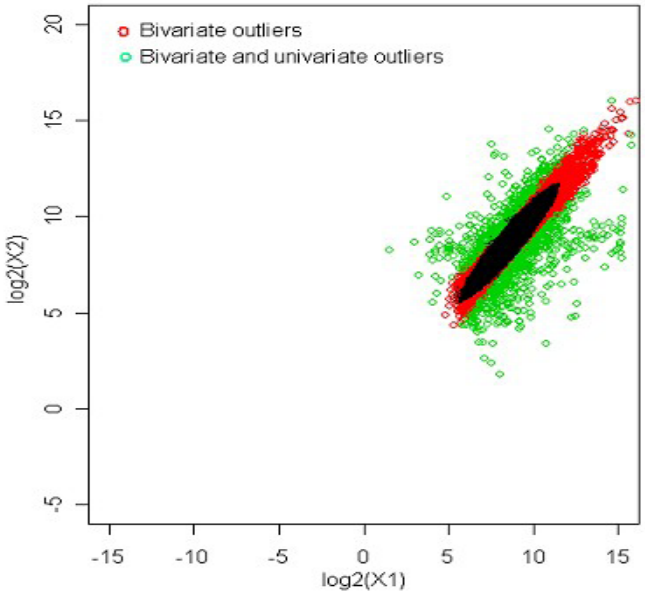
Mehalanobis Distance:

The covariance matrix is used for the quantification of the size and shape of the multivariate data, which is taken into account in the Mahalanobis Distance. For a multivariate sample Xij, where i = 1,2,3,...n (number of genes) and j =1,2,3...p (number of samples), the Mahalanobis distance is defined as,

MDi=((Xij– m)T C-1(Xij - m))0.5

where m is estimated multivariate location parameter and C is the estimated covariance matrix. The location and the covariance parameters are determined using Minimum Covariance Determinant estimation method. The MCD estimator is determined by that subset of observations of size *h*, which minimizes the determinant of the covariance matrix computed only

from the *h* observations. The location estimator is the average of these *h* observations, whereas the scatter estimate is proportional to the variance covariance matrix.



. *Fig 5.3 Outliers obtained using bivariate and univariate approaches.*

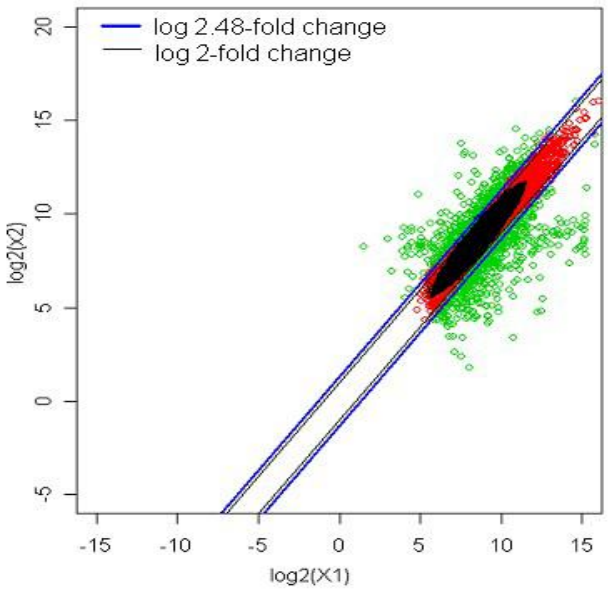


Fig 5.4 The up and down regulated genes for 2.48- and 2- log fold change

Stage-II: Univariate Outlier detection:

Let S denote the original set of observations. Let Sout and Sin be the subsets of S containing outlier and inlier observations respectively. Thus, SoutU Sin = S andSout∩Sin = {Ø}, i.e. the two subsets are mutually exclusive.

We denote:

Sout = {(log 2(Xi1), log 2(Xi2)) / MDi> c for i=1,2,3...n} and

Sin = {(log 2(Xi1), log 2(Xi2)) / MDi ≤ c for i=1,2,3...n}

where 'c' is the cut-off for a given quantile and n is the total number of genes.

We define a statistic:

Z = log2 (X2 / X1) = log2(X2) – log2(X1)

which is the log of the ratio of intensity values for different genes for the two samples.

In the current study, Gene Expression Analysis was performed to find out differentially expressed genes between Type-2 Diabetes with and without parental history. For this analysis Multivariate and Univariate outlier detection methods are used. This analysis helps in identifying the potential Candidate Genes causing Type-2 Diabetes. Out of 3940 outlier genes, 1211 were detected as up-regulated, while 368 were detected as down-regulated genes with respect to the healthy (H) individual as a case study . Thus, for healthy vs. diabetic with parental history comparison, 1579 genes were found to be differentially expressed out of 39400, which amounts to 4% of the total genes under study. This is 2.73% less than the number of genes obtained for 2-fold change thresholds (Table 5.1).

S.No.	Reference sample	Test Sample	% of DE genes for 2 fold change	Modified fold change (MFC)	% of DE genes for MFC	Up-Regulated Genes	Down-Regulated Genes
1	H	D&PH	6.73	2.36	3.33	1211	368
2	H	D&NPH1	7.69	2.37	4.38	1249	477

Table 5.1 : Sample results showing Up and Down regulated genes

6. CLASSIFICATION OF IDENTIFIED GENES

The identified differentially expressed genes are further functionally classified using Gene Ontology and Pathway analysis.

Gene Onlology (GO) :

Gene Ontology (GO) is a major public annotation effort, which provides descriptions of the molecular functions, biological processes and sub-cellular locations attributed to gene products from all organisms. GO links primary biological knowledge to information provided in highly-controlled, structured vocabularies (or ontologies), and is designed to improve the accessibility of scientific knowledge to search engines and algorithmic processing. Consequently, GO has

proved to be highly beneficial to investigators who need to understand and analyze large amounts of data produced from a range of high-throughput investigative techniques.

There is no universal standard terminology in biology and related domains, and term usages may be specific to a species, research area or even a particular research group. This makes communication and sharing of data more difficult. The Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains:

- **Molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis. In the current study, genes involved in NADH dehydrogenase (ubiquinone) activity, glutamate dehydrogenase [NAD(P)+] activity, CDP-diacylglycerol-glycerol-3-phosphate-3-phosphtidyltransferase activity are upregulated in D&PH with respect to H. Gene involved in protein kinase B binding, enzyme inhibitor activity, acyl-CoA oxidase activity, phosphatidylinositol transporter activity, acyltransferase activity are down regulated in D&PH with respect to H.

- **Biological Process** , operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. In the present study, genes involved in synaptic vesicle membrane organization and biogenesis, polysaccharide metabolic process, regulation of growth rate, nucleosome assembly are up regulated in D&PH with respect to H. Genes involved in immune response, regulation of glycolysis are down regulated in D&PH with respect to H.

- **Cellular Component**, the parts of a cell or its extracellular environment; Genes localized in cohesin core heterodimer, oligosaccharyltransferase complex, nucleosome, respiratory chain complex II are up regulated in D&PH with respect to H. Genes localized in isoamylase complex, protein kinase CK2 complex, proteasome activator complex, 6-phosphofructokinase complex are down regulated in D&PH with respect to H.

Pathway Analysis :

The development of microarray technology allows the simultaneous measurement of the expression of many thousands of genes. The information gained offers an unprecedented opportunity to fully characterize biological processes. However, this challenge will only be successful if new tools for the efficient integration and interpretation of large datasets are available. One of these tools, pathway analysis, involves looking for consistent but subtle changes in gene expression by incorporating either pathway or functional annotations. Pathway analysis is a promising tool to identify the mechanisms that underlie diseases, adaptive physiological compensatory responses and new avenues for investigation. Pathways are collections of genes and proteins that perform a well-defined biological task.

In the present study, genes involved in Inositol phosphate metabolism, Starch and sucrose metabolism, Nitrogen metabolism, Oxidative phosphorylation, Androgen and estrogen metabolism, Glycan biosynthesis and metabolism pathways, Metabolism of cofactors and vitamins pathways, MAPK signalling pathway, ECM-receptor interaction, Neuroactive ligand-receptor interaction, Regulation of actin cytoskeleton, Cell communication pathways, Nervous system pathways, Neurodegenerative disorders pathways are up regulated in D&PH Vs H. Genes involved in Glycolysis / Gluconeogenesis, Propanoate metabolism, Carbon fixation, Biosynthesis of steroids, Fatty acid metabolism, Histidine metabolism, Phenylalanine metabolism, Tyrosine metabolism, Urea cycle and metabolism of amino groups, Cell cycle, Insulin signalling pathway, PPAR signaling pathway, Antigen processing and presentation are down regulated in D&PH Vs H.

Condition	Differentially expressed genes concerned with inflammation
Diabetic with family history vs healthy individual (DPH vs H)	ALK, GCH1, IFIH1, IFIT1, ILIIRA, ITGB2, MAP3K4, MMP19, MMP3, RPS27A, SLK, TNFRSF12A, UBC
Diabetic with family history vs Diabetic without family History (DPH vs DNPH1)	ALK, CCL13, CCR8, CDKNIA, EDN1, FGF1, IFIT1, ILI2RB1, IL20, IL22, IL2RG, IL8RA,ITGB2,MMP20,SLK,TNFTFS12A,UBC, XCR1
Diabetic with family history vs Diabetic without family History (DPH vs DNPH1)	ALK, BLRI,CCL15, CCL16, CCR7, CCR8, CXCL11, CXCL12, FN1, FTH1, GBP1, HLA-A, IFIT1, IL 12A, ITGB2, KIT, LTB, MMP20, PPARD, RHOA, RPS27A, TAC1, TLR4, TNFAIP6, TNFRSF11A

Table 6.1: Genes involved in inflammatory response that were differentially expressed in Type-2 diabetes mellitus with family history for various test cases.

7.CONCLUSION

The incidence of type 2 diabetes mellitus and parental history is the main focus in the present work. This is supported by the results of the present study wherein it was noted that genes involved in inflammatory response are differentially expressed in subjects type 2 diabetes mellitus with parental history vs healthy and without parental history (Table 6.1). These results suggest that low-grade systemic inflammation plays a significant role in the pathobiology of type 2 diabetes mellitus, and parental history. Thus, the results of the present study and other investigations indicate the genes concerned with parental history and healthy response are differentially regulated in type 2 diabetes mellitus. The present results need to be verified by estimating the concentrations of the specific proteins of the genes expressed, and studying more closely the interaction(s) between the nervous system, hypothalamic peptides and neurotransmitters, pro and anti-inflammatory cytokines, and their relationship to appetite, satiety and development of type 2 diabetes mellitus.

8.ORGANISATION OF THE THESIS

In Chapter 1, issues related to Type 2 diabetes mellitus, use of bioinformatics, microarray technology, Gene expression analysis and normalization of microarray data including the recent trends are introduced.

Chapter 2 elaborated the review on different methods of analyzing gene expression data. A detailed review on various methods of microarray data analysis and gene expression analysis is presented. Discussed the merits and demerits of various methods and described the statement of problem and methodology of current study.

Chapter 3 introduces the differential expression analysis of Type 2 diabetes mellitus with parental history. Outlier approach using Mahalanobis Distance measure is discussed in detail. Analysis between combination of samples from three categories viz. Healthy, Diabetic with parental history and Diabetic without parental History are presented.

Chapter 4 describes the functional classification of differentially expressed genes. Describes the use of Gene ontology analysis process to know the cellular component, molecular function and biological process of the differentially expressed genes. Discusses pathway analysis process and presented the results from various combinations of test samples.

Chapter 5 provides a detailed summary of the work with salient features. This chapter explores open problems for future enhancements.

9.REFERENCES

[ADI, 2006] Adi L. Tarca, Roberto Romero, Sorin Draghici, 'Analysis of microarray experiments of gene expression profiling' American Journal of Obstetrics and Gynecology 195, pp 373–88, 2006.

[ANJ, 2006] Anja Schienkiewitz, Matthias B Schulze, Kurt Hoffmann, AnjaKroke and Heiner Boeing, 'Body mass index history and risk of type 2 diabetes: results from the European

Prospective Investigation into Cancer and Nutrition (EPIC)–Potsdam Study', Am J Clin Nutr 2006;84:427–433, 2006.

[ATH, 2009] Athanasia Papazafiropoulou, AlexiosSotiropoulos, EystathiosSkliros Marina Kardara, AnthiKokolaki, OuraniaApostolou and Stavros Pappas, 'Familial history of diabetes and clinical characteristics in Greek subjects with type 2 diabetes', *BMC Endocrine Disorders*, pp **9**:12, 2009.

[CHE, 1997] Chen, Y., Dougherty, E. R., Bittner, M. L. 'Ratio-based decisions and the quantitative analysis of cDNA microarray images' , J. Biomed. Optics 2(4):364–374 1997.

[DEL, 2002] Delongchamp, R. R., Velasco, C., Evans, R., Harris, A., Casciano, D, 'Adjusting cDNA Array for Nuisance Effects' , Technical Report, Jefferson, AR: National Center for Toxicological Research, 2002.

[EFR, 2000] Efron, B., Tibshirani, R., Goss, V., Chu, G. 'Microarrays and Their Use in a Comparative Experiment' Preprint 37B/213. Stanford University, 2000.

[FAT, 2004] Fatima Sanchez-Cabo, Andreas Prokesch, Gerhard G. Thallinger, Roland Pieler and Zlatko Trajanoski, Philip D. Butcher, Jason Hinds, Leah E. A. Holmes, Susan G. Campbell, Mark P. Ashe, Simon Hubbard, Kwang-Hyun Cho, Olaf Wolkenhauer, 'Assessing the efficiency of dye-swap normalization to remove systematic bias from two-color microarray data' CiteseerX, 2004.

[FRA, 2010] France Pouwer, 'Nina kupper, Marcel C Adriaanse, 'Does emotional stress cause Type-2 Diabetes Mellitus ? A Review from the European Depression in Diabetes (EDID) Research Consortium'**,** Discovery Medicine, 9(45), 112:118, 2010.

[HAI, 2004] Haifeng Li , Keshu Zhang , and Tao Jiang, ' Minimum Entropy Clustering and Applications to Gene Expression Analysis', CSB 2004 Proceedings, 2004 IEEE (Aug 2004), pp. 142-151, 2004.

[HAN, 2011] Hannah Bruehl , Victoria Sweat , Aziz Tirsi1, Bina Shah , Antonio Convit, 'Obese Adolescents with Type 2 Diabetes Mellitus Have Hippocampal and Frontal Lobe Volume Reductions' *Neuroscience & Medicine*, Vol 2, PP 34:42, 2011.

[Karter AJ et.. al 1999] Karter AJ, Rowell SE, Ackerson LM, Mitchell BD, Ferrara A, Selby JV, Newman B, 'Excess maternal transmission of type 2 diabetes: the Northern California KaiserPermanente Diabetes Registry' .*Diabetes Care* 22,938:943, 1999.

[KER and CHU, 2001] Kerr, M. K., Churchill, G. A. 'Experimental design for gene expression Microarrays',. Bio-Stat. 2:183–201, 2001.

[KER, 2001] Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., Churchill, G. A, 'Statistical analysis of a gene expression microarray experiment with Replication', Stat. Sinica 7(6):819–838, 2001.

[KLE, 1996] Klein BE, Klein R, Moss SE, Cruickshanks KJ 'Parental history of diabetes in a population- based study', *Diabetes Care* 19: 827–830, 1996.

[LOU, 2007] Louise a. Kelly, christianne j. Lane, marc j. Weigensberg, corinna koebnick, christian k. Roberts, jaimie n. Davis, claudia m. Toledo-corral, gabriel q. Shaibi michael i. Goran, 'Parental History and Risk of Type 2 Diabetes in Overweight Latino Adolescents - A longitudinal analysis', Diabetes care, volume 30, number 10, 2700:2705, 2007.

[MAY, 2006] Mayor S 'Diabetes affects nearly 6% of the world's adult'**.** *Br Med J* 333**:**1191, 2006.

[PAB, 1999] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, SutisakK kitareewan, Ethan Dmitrovsky, Eric s. Lander, and Todd r. Golub, ' Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation', Genetics, Vol. 96, pp. 2907–2912, 1999.

[SAT and YAS, 2009] Satoshi Niijima and Yasushi Okuno, 'Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection' IEEE/ACM transactions on computational biology and bioinformatics, vol. 6, no. 4, 605:614, 2009.

[SCH, 2000] Schuchhardt, S., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H. 'Normalization strategies for cDNA microarrays' , Nucleic Acids Res. 28(10):e47, 2000.

[SPE, 1998] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B, 'Comprehensive identification of cell cycleregulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization', Mol. Biol. Cell 9(12):3273–3297, 1998.

[WOL, 2001] Wolfinger, R. D., Gibson, G., Wolfinger, E. D, 'Assessing gene significance from cDNA microarray expression data via mixed models' Journal Of Computational Biology, Volume 8, Number 6, Mary Ann Liebert, Inc. Pp. 625–637, 2001.

[XIA and ILL, 2004] Xiaohua Hu, Illhoi Yoo 'Cluster Ensemble and Its Applications in Gene Expression Analysis',APBC '04 Proceedings of the second conference on Asia-Pacific bioinformatics - Volume 29, 2004.

LIST OF PUBLICATIONS FROM THE WORK

1. Chandra Sekhar Vasamsetty, , Srinivasa Rao Peri, Allam Appa Rao, K. Srinivas, and Chinta Someswararao, 'Gene Expression Analysis for Type-2 Diabetes Mellitus –A Case Study on Healthy vs Diabetes with Parental History', IACSIT International Journal of Engineering and Technology, Vol.3, No.3, pp 310-314,2011.

2. Chandra Sekhar Vasamsetty, Dr. Srinivasa Rao Peri, Dr. Allam Appa Rao, Dr. K. Srinivas, Chinta Someswararao, 'Gene Expression Analysis for Type-2 diabetes mellitus – A study on diabetes with and without parental history' Journal of Theoretical and Applied Information Technology, Vol.27, No.1, pp 43-53, 2011.

3. V Chandra Sekhar, Allam Appa Rao, P.Srinivasa Rao, K.Srinivas, 'Identification of differentially expressed genes for diabetes with parental history vs healthy using Microarray data analysis' IEEE 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Vol:4, pp 496-500, 2010.

4. V Chandra Sekhar, Allam Apparao, P.Srinivasa Rao , 'Differential Gene Expression Analysis for Diabetes with and without Parental History', 3rd IEEE International Conference on Computer Science and Information Technology, Vol:9, pp 322-326, 2010.